Week 1 : Introduction

pending tasks

type

What is data science?

There is a lot of confusion around the term since it is being used as a buzzword also, the field encompassed is large.

The following are the tasks in a data science pipeline:

- collect
- store
- proecss
- describe
- model

Collecting Data

def data science (t = 0:20)

The task depends on the objective and working environment (ie. if the org. is data rich). It involves sourcing required structured and unstructured data from respective sources or designing experiments to collect the required data.

Collection of data from structured sources requires knowledge of databases. To pull data using APIs or scraping from the internet requires intermediate programming skills. Designing experiments to collect the required data calls for understanding of statistics(addressing bias and variable dependencies).

Skills: understanding DB, statistics and programming skills.

Storing Data

def Big Data (t = 7:49)

Relational databases - To store and update data. Data is structured and optimized for SQL queries.

Data warehouses - To integrate structured data from multiple repositories and optimized for analytics.

Data lake - Collection of all data of all formats related to an org. also, the data is not curated.

Big Data is characterized by volume, variety and velocity.

Skills : Programming , Relational DB, NoSQL DB, data warehouses, data lakes (Hadoop).

Processing Data

1. Extraction, transformation and loading form the process of data wrangling/munging. It involves collecting and converting data to the required format(with respect to storage).
2. (t = 4:14) Data cleaning handles missing values, standardisation of keywords, addressing spelling errors and removal of outliers.
3. (t = 6:29) Data scaling, normalisation and standardization

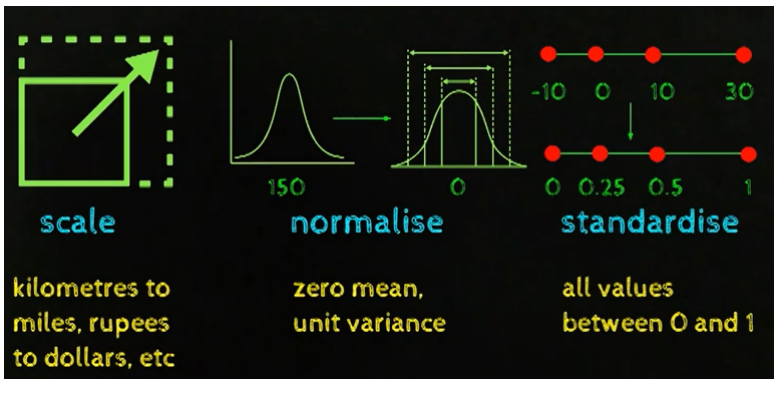

Fig.1 Data scaling, normalization and standardization.

Processing big data involves processing data in chunks, this can be distributed using frameworks such as Hadoop(Map Reduce).

Skills : Programming, Map Reduce, SQL and NoSQL, statistics.

Describing Data

It comprises two aspects, data visualization and data summary.

Data visualization : using different plots a sense of data characteristics can be quickly drawn.

(t = 3:43) Summarizing data : summary statistics typically involve computation of mean, median, mode, standard deviation and variance from the data.

Descriptive statistics is an iterative process also known as Exploratory Data Analysis (EDA).

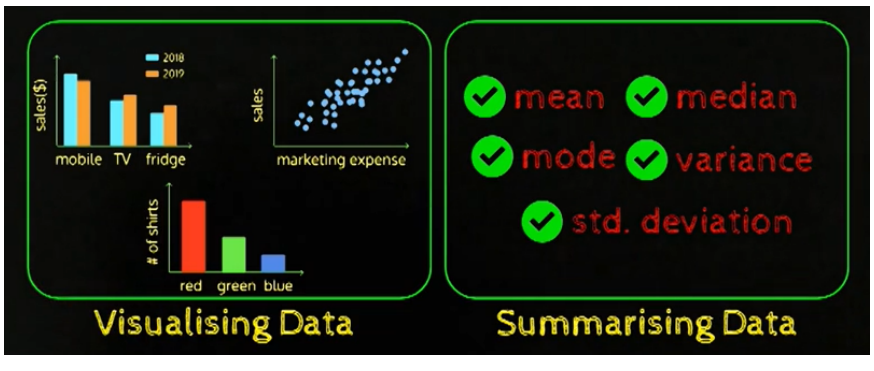

Fig.3 Data visualization and summary

Skills : Statistics, Excel, Python, R, Tableau. To draw Business Intelligence(BI) either BI tools or programming languages can be used.

Modeling : statistical and Algorithmic

Modeling can be statistical or algorithmic depending on the requirement.

Statistical modeling (video 1): Statistics is used to understand the underlying data distribution.It helps make robust arguments using the data available.

Statistical modeling helps to discover underlying relationships between the data attributes (variables), formulating and testing hypotheses and, to give statistical guarantees.It is more suitable for low dimensional data.

Algorithmic modeling (video 2): Used when the relationship between input and the output is complex. The function is estimated using the data and optimisation techniques. The goal is to get the output not how it is obtained ie. on prediction.The data used can be high dimensional.

Statistical modelling vs. Algorithmic modeling (t = 6:17)

Why is data science so popular today?

Data is being generated and consumed in large amounts. Also, hardware required to support data science has become cheaper and more efficient. Software and hardware resources have been democratized in terms of open source software and cloud compute respectively.

Are AI and data science related?

def AI (t = 2:23)

The terms have been broadly used in the mainstream media, thus, leading to misleading images of what they technically stand for. AI is about building systems or agents that demonstrate intelligence. It constitutes the tasks of problem solving, knowledge representation, reasoning, decision making and perception, communication and actuation.

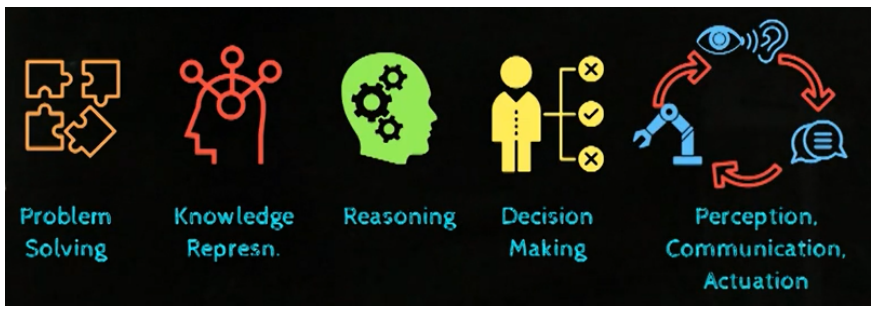

Fig.4 Tasks in AI pipeline.

Decision making, perception, communication and actuation are data driven and intersect with the field of data science.

Problem Solving

Using efficient search algorithms, the next best action is selected.

Knowledge Representation and Reasoning

Propositional logic is used for knowledge representation. Reasoning is done by executing inference rules(sequential decision making based on knowledge). These are not data data driven. It only requires understanding of propositional and first order logic.

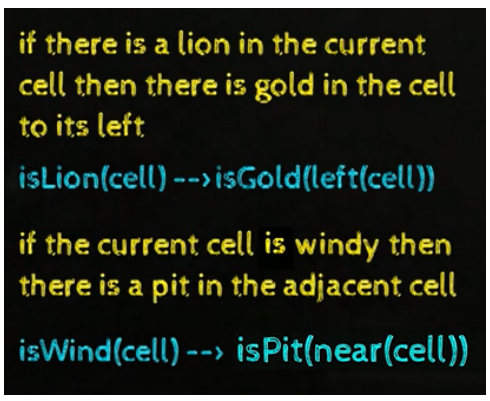

Fig.5 Example of first order logic.

TEXTBOOK for AI : Artificial Intelligence: A Modern Approach (Stuart J. Russell and Peter Norvig)

Decision Making

Expert systems are used for decision making. The rules are given by the domain experts and encoded using knowledge representation. The reasoning and rule execution are done by a program.

Expert systems cannot be implemented when the rules are too complex, inexpressible or unknown. Thus, an alternative approach of arriving at the output using large amounts of data ie. data science is used. The decision making can be done using ML, DL or RL algorithms.

Reinforcement learning: the agent is in a dynamic environment and has to make sequential decisions. There are states and defined actions and no explicit supervision at each step. The agent is offered rewards from the environment as a reinforcement to each decision taken, this ought to drive it to take next best action.

Communication, Perception and Actuation

def Data Mining : (t = 10:49)

This task also intersects with data science with the usage of NLP and CV for communication and perception respectively. NLP and CV largely use DL and are data driven. Speech recognition also falls under the category of perception and uses DL techniques.

Actuation is done using physical robots. The robotic actuation is now data driven, learnt from simulation or by mimicking human examples.

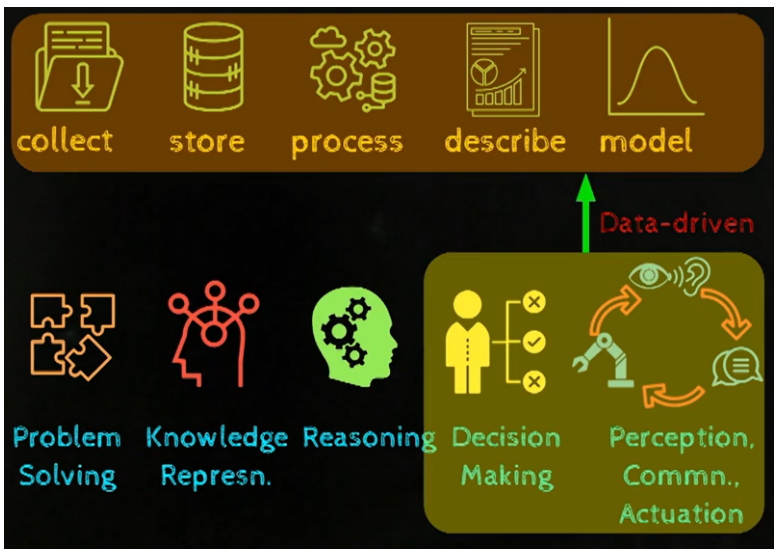

Fig.6 Intersection of AI and data science.

Data mining is performing data science on structured data.

The myths of data science.

Machines cannot make decisions on their own. Human intervention and insights are required in each step of the data science pipeline. Machines only execute the script and provide the required hardware to do so efficiently. Any intelligence exhibited by a machine is programmed by a human.

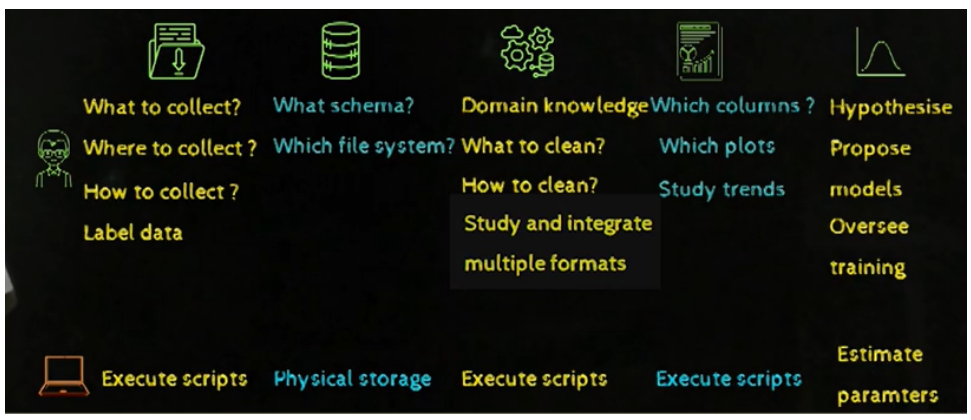

Fig.7 Roles of human vs computer at each stage of pipeline.

Data science also includes statistical learning. It is not limited to big data and DL.

Data science is not always successful. There may not be meaningful or actionable insights present in the data. DS may succeed when a large amount of clean data is provided.

Skilled data scientists and resources are required to make the output impactful.

The path to data science.

Understanding of programming, databases and mathematics are essential foundations of data science.

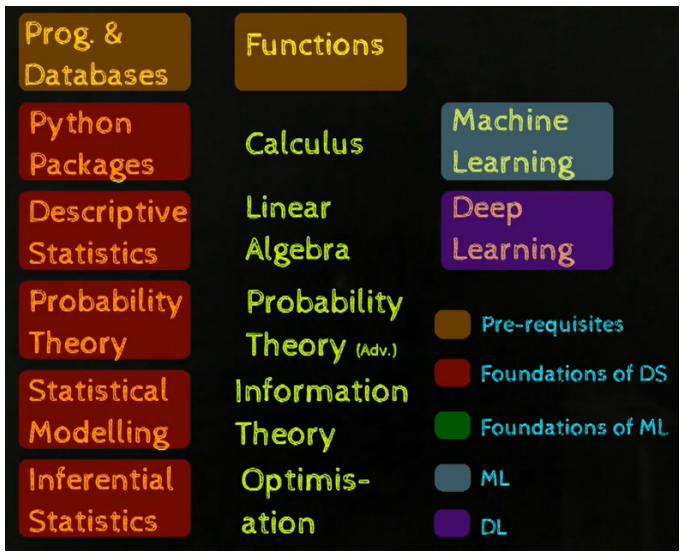

Fig.8 Skills required for data science.

Foundation of ML required understanding of various algorithms used, the math involved and optimization techniques.

Summary

- A typical data science pipeline involves the five stages of data collection, storage, processing, describing amd modeling.
- The popularity of data science is due to generation of large volumes of data and availability of hw/sw platforms to process them.
- The fields of AI and data science intersect when the decisions made by the expert systems are data driven rather than encoded rules given by the domain expert.
- Each stage in the pipeline of data science involves both human and machine.
- DL is a subset of data science. Also not all activities are successful.
- Learning data science requires strong foundations in mathematics and intermediate programming skills. Any approach excluding either will not give a holistic view of the subject.

MCQ

1. The company X has taken up a new project that requires creation of a data management platform (DMP). This involves collecting data from various sources and integrating them to create a single view of all sources. What are the steps involved in this process?
    1. data collection, storage and processing
    2. Data collection, ETL, modeling
    3. **Data collection, wrangling and QC, describing**
    4. None of the above

2. In a given dataset more than 30% values are missing. How can this be handled?
    1. **Remove rows with missing values if the dataset is large.**
    2. Fill the missing values with value from the previous record.
    3. **Fill missing values using mean or average of the remaining records.**
    4. **Fill the value using another record with similar attribute values.**

3. Data science can be described as
    1. **Field of study used to validate hypotheses or derive insights from data.**
    2. **Set of tasks that aid in making data-driven decisions.**
    3. Making sense of data using ML techniques
    4. All of the above

4. Data science is enabled by
    1. Skilled data scientists and engineers.
    2. Availability of data, software and hardware resources.
    3. The need to aid decisions made based on numbers.
    4. **All of the above.**

5. The company X has taken up a new project that requires creation of a data management platform (DMP) of its retail stores. This involves collecting data from various sources and integrating them to create a single view of all sources. How is the data stored?
    1. A relational database is used to store the transformed data from different sources.
    2. A data lake is created.
    3. **Data is assimilated into a data warehouse and refreshed daily.**
    4. Any of the above can be used.

6. Examples of AI are
    1. **A robot that collects cans from the room.**
    2. **Air conditioning with automatic temperature adjustments.**
    3. A script that automatically downloads content from a webpage.
    4. Only the robots that portray human characteristics such as perception and reflex.

7. With the invent of autoML and other user friendly frameworks, it is not necessary to deep dive into the nuts and bolts of machine learning ie. mathematical nuances.
    1. True
    2. **False**

8. AI is a subset of data science.
    1. True
    2. **False**

9. A chess playing robot uses
    1. ML
    2. DL
    3. **RL**
    4. Statistical analysis of it’s position.

10. Machine learning implies
    1. **When given input and output, the machine finds out the most suitable function that can map the input to the output.**
    2. When given input and program, the machine finds out the output.
    3. **Estimation of parameters using machines.**
    4. Explanation for the relationship between the input and output variables is given by the machine.